VOCALIST GENDER RECOGNITION IN RECORDED POPULAR MUSIC

Björn Schuller, Christoph Kozielski, Felix Weninger, Florian Eyben and Gerhard Rigoll
Institute for Human-Machine Communication
Technische Universität München
Munich, Germany
{lastname}@tum.de

ABSTRACT

We introduce the task of vocalist gender recognition in popular music and evaluate the benefit of Non-Negative Matrix Factorization based enhancement of melodic components to this aim. The underlying automatic separation of drum beats is described in detail, and the obtained significant gain by its use is verified in extensive test-runs on a novel database of 1.5 days of MP3 coded popular songs based on transcriptions of the Karaoke-game UltraStar. As classifiers serve Support Vector Machines and Hidden Naive Bayes. Overall, the suggested methods lead to fully automatic recognition of the pre-dominant vocalist gender at 87.31 % accuracy on song level for artists unkown to the system in originally recorded music.

1. INTRODUCTION

Determination of the gender of the (main) vocalist(s) is an astonishingly untouched task in the field of Music Information Retrieval (MIR): while there is a substantial body of literature dealing with gender in spoken language, e. g. to improve automatic speech recognition systems by switching or adapting acoustic models (e. g. [1]) or accordingly to improve emotion recognition systems (e. g. [22]), only some works consider singer identification (on artificial signals) [3, 10]. However, explicit recognition of the gender of the main performing vocal artist in original audio recordings of e. g. contemporary popular music has apparently not been addressed in MIR research, yet, which is to overcome, as like genre, mood or style, it can be an important feature for organizing and querying music collections, for example to find a song whose artist's name is unknown to the user, or for recommendation systems in on-line stores. In addition, it might be considered interesting as mid-level attribute for other MIR tasks as audio mood classification [14] or transcription of the sung lyrics with gender-adapted models – shown to be beneficial in [9].

Apart from finding suitable features and an appropriate classification method, as is the pre-concern for gender identification in (spoken) speech analysis, a setting dealing with the named original audio recordings of music demands for reasonable enhancement of the singer(s) voice given the background 'noise' of musical and rhythmic accompaniment. It is comparably easy to *eliminate* the main singer's voice in stereophonic recordings, e. g. for Karaoke application: often stereophonic channel subtraction killing the mid-panned parts suffices, as the lead singer's voice is usually panned there to be well audible at any position carrying the main melody (in fact the bass is usually panned there as well, which can be by-passed first). However, to *separate* these vocals is a non-trivial and challenging task – 'intelligent', i. e. data-driven, spectral decomposition is usually required to this end.

Non-negative Matrix Factorization (NMF) is one of the increasingly popular algorithms used within blind source separation of audio signals. Among other fields (e. g. speech recognition [18] or non-linguistic vocalisation recognition [16]), it has been successfully used in speaker separation [12], instrument separation [17, 23], especially drum beat separation [4, 15, 20], and vocal separation [10, 21]. While these methods provide audible results of great quality, it is not fully clear to which extent blind source separation can aid in general Music Information Retrieval tasks [15]. Here, we employ it to separate drum-beats from the rest of a popular music piece. While one could directly aim at separation of the vocals, this is considerably more difficult giving the large spectral overlap with other instruments. We thus decided to remove the relatively easier separable drum and percussion part and recognize gender in combination with general vocal presence in the remaining audio.

In this work the recognition system has to identify the gender of the performing artist particularily on 'real-world' data, i. e. originally recorded music without any preselection of 'friendly cases', which is a challenging task not only due to the above-named instrumental accompaniment, but also to the variety of genre and singing styles. In our experiments we introduce a database of popular songs from the Karaoke game UltraStar, which includes annotations of the tempo and the location of the sung parts. Gender is additionally labelled for the following experiments.

In the remainder of this paper, we first introduce the UltraStar database in section 2, then explain the acoustic features and the classifiers used for vocalist gender recognition in music in section 3, and the methodology applied for separating the drum beat with NMF in section 4. Then, section 5 presents the data partitioning throughout test-runs and the experimental results before finally deriving conclusions in section 6.

2. ULTRASTAR DATABASE

We first introduce a data set of songs with annotations in the file format of the open-source Karaoke game UltraStar [2] for vocalist gender evaluation, referred to as *UltraStar database* in the ongoing. This set contains 582 complete songs and is encoded in MP3 (MPEG-1 Audio Layer 3) format with 44.1 kHz PCM and variable bit rate with a minimum of 128 kbit/s. In total length, this set corresponds to 37 h 06 min of music, i. e. 1.5 days of continuous music. The set covers generations from the 1960s until today and is a good example of typical popular music from diverse genres like Rock, Pop, Electronic Music, Ballads or Musical.

As annotation, we use the tempo and the information on the location of vocal presence in a song. In addition, we carried out a gender annotation on song level for this data set: per song we assigned the gender of the vocalist that is perceived as pre-dominant over the total run-time after listening to each full song. This was done by two labellers individually and in random order without any disagreement. Overall, 178 songs were labelled as female, and 404 songs as male, respectively 11 h 08 min female and 25 h 58 min male playtime. Prior to the processing all songs were down-mixed to monophonic by non-clipping stereo channel-addition.

Since every user of the UltraStar Karaoke game has the possibility to contribute annotations to the game's website, we chose songs according to their popularity among users, assuming that high popularity of a song indicates that a robust ground truth can be established.

3. GENDER RECOGNITION

3.1 Acoustic Features

For the actual gender recognition we consider the short-time energy, zero-, and mean-crossing rate known to indicate vocal presence [24]. In addition we extract values from the normalized autocorrelation sequence of the DFT coefficients, namely voicing probability, F-zero, and harmonics-to-noise ratio (HNR). F-zero is the location of the highest peak of the autocorrelation sequence aside from the maximum at zero. HNR is computed by the value of this peak. We further calculate Mel frequency cepstral coefficients (MFCC) 0–13 and their respective first-order delta regression coefficients. MFCC are known to capture the characteristic qualities of individual voices in speech and music for singer identification [8, 10, 11] and have proven highly meaningful in various speech gender recognition tasks [1, 22]. Thus, altogether we employ a set of 32 features.

Vocals in popular music are synchronous to the beats of a song most of the time. For every quarter beat we know from the annotations in the UltraStar database whether sung vocals are present or not. From this we derived an annotation on beat level by a non-ambiguous majority vote procedure: we judged vocals to be present in a beat if they are present in at least two of the quarter beats.

Based on this, every song in our data set is divided into analysis frames corresponding to the beats of the song. As the tempo and the locations of vocal presence are known for each song, beat synchronous chopping is possible for the training section and test section, to focus on the problem at hand. However, using the highly reliable automatic beat-tracker as introduced in [13] led to non-significant (one-tailed test, for testing conditions cf. below) differences in accuracy on song level. We divide the signal into non-overlapping frames with a Hamming window function of the length of a beat of the particular song – a strategy found beneficial over smaller units in previous tests. Per likewise beat-synchronous frame the above mentioned features are computed.

For easy reproducibility of the results we decided for open-source feature extraction by using the real-time toolkit for 'Speech and Music Interpretation by Large Space Extraction' (openSMILE) [1] .

3.2 Classifiers

We evaluate Support Vector Machines (SVM) with polynomial Kernel, sequential minimal optimization learning, and pairwise multi-class decision, as well as different Bayesian classifiers for our gender recognition task to be more independent of classifier influence.

A Bayesian network in general is a directed graph in which nodes represent attributes and branches represent attribute dependencies. It is quantified by conditional probabilities for each node dependent on its parents. In naive Bayes, each attribute node has the class node as its parent only, without any relation to other attributes, so it is the simplest Bayesian network [6]. In structure learned Bayesian networks every attribute node can have other attribute nodes as its parents, thus all dependencies between attributes are considered. However, achieving an optimal structure by learning from data is often impracticable in reasonable time. Hidden naive Bayes represents attribute dependencies by creating a hidden parent for each attribute node. This parent combines all influences of other attributes. Although attribute dependencies are considered, it keeps the structure of naive Bayes and does not need to be structure learnt [25].

For Bayes classification we found discretization by Kononenko's minimal description length criterion [5] based on the training instances beneficial (significant gain in average accuracy of 5.54 % on beat level with and without enhancement in the experiments as follows, for testing conditions cf. below), and Hidden Naive Bayes (HNB) superior to the considered alternatives (significant gain as before of 2.41 % over structure learned Bayesian networks, and of 7.68 % over Naive Bayes).

[1] http://www.openaudio.eu

4. DRUM BEAT SEPARATION USING NON-NEGATIVE MATRIX FACTORIZATION

4.1 Definition of NMF

Given a matrix $\mathbf{V} \in \mathbb{R}_{\geq 0}^{m \times n}$ and a constant $r \in \mathbb{N}$, non-negative matrix factorization (NMF) computes two matrices $\mathbf{W} \in \mathbb{R}_{\geq 0}^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times n}$, such that

$$\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H} \tag{1}$$

For information reduction one generally chooses r such that $(m + n)r \ll mn$.

4.2 Application to Blind Source Separation

An important application area of NMF in signal processing is blind source separation. In the particular field of music processing, NMF has been successfully used to separate drum from harmonic sounds [4, 15, 20].

NMF-based blind source separation is usually realized in the frequency domain. Thereby the signal is split into overlapping frames of constant size. In our experiments, a frame size of 60 ms and an overlap of 50 % produced best results. Each frame is multiplied by a window function and transformed to the frequency domain using Discrete Fourier Transformation (DFT), with transformation size equal to the number of samples in each frame. We use the square root of the Hann function for windowing, as this helps to reduce artifacts when transforming back to the time domain [4].

Only the magnitudes of the DFT coefficients are retained, and the frame spectra are put in the columns of a matrix. Denoting the number of frames by n and the frame size by T, and considering the symmetry of the coefficients, this yields a $(\lfloor T/2 \rfloor + 1) \times n$ real matrix.

To exploit NMF for blind source separation, one assumes a *linear signal model.* Note that Eq. 1 can be written as follows (the subscripts $:, t$ and $:, j$ denote the t^{th} and j^{th} matrix columns, respectively):

$$\mathbf{V}_{:,t} \approx \sum_{j=1}^{r} \mathbf{H}_{j,t} \mathbf{W}_{:,j}, \quad 1 \leq t \leq n \tag{2}$$

Thus, if $\mathbf{V}$ is the magnitude spectrogram of a signal (with short-time spectra in columns), the factorization from Eq. 1 represents each short-time spectrum $\mathbf{V}_{:,t}$ as a linear combination of spectral basis vectors $\mathbf{W}_{:,j}$ with non-negative coefficients $\mathbf{H}_{j,t}$ $(1 \leq j \leq r)$.

We define the j^{th} *component* of the signal to be the pair $(\mathbf{w}_j, \mathbf{h}_j)$ of a spectrum $\mathbf{w}_j := \mathbf{W}_{:,j}$ along with its time-varying gains $\mathbf{h}_j := \mathbf{H}_{j,:}$ (the subscript $j, :$ denotes the j^{th} matrix row).

It has turned out that the non-negativity constraint on the coefficients alone is sufficient to decompose a signal into the underlying sources [17, 20]. Note that a 'source' in the intuitive sense, e. g. an instrument, can consist of multiple components.

When there is no prior knowledge about the number of spectra that can describe the source signal, the number of components r has to be chosen empirically. In our experiments, best results were achieved by setting $r = 30$.

4.3 Factorization Algorithm

A factorization according to Eq. 1 is usually achieved by iterative minimization of cost functions. For the purpose of drum beat separation, an extended form of the Kullback-Leibler (KL) divergence has been shown to yield good results [15, 20]:

$$c_d(\mathbf{W}, \mathbf{H}) = \sum_{i=1}^{m} \sum_{t=1}^{n} \left(\mathbf{V}_{i,t} \log \frac{\mathbf{V}_{i,t}}{(\mathbf{WH})_{i,t}} - (\mathbf{V} - \mathbf{WH})_{i,t} \right) \tag{3}$$

Eq. 3 can be enhanced by a term that enforces temporal continuity of the gains, improving separation quality [20] at the expense of increased computational costs. Because the perceived audio quality of components is of minor relevance for our task, we chose c_d from Eq. 3 as cost function and minimized it using Lee and Seung's multiplicative update algorithm [7]. It performs the following iterative updates of the matrices $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{H}_{j,t} \leftarrow \mathbf{H}_{j,t} \frac{\sum_{i=1}^{m} \mathbf{W}_{i,j} \mathbf{V}_{i,t} / (\mathbf{WH})_{it}}{\sum_{i=1}^{m} \mathbf{W}_{i,j}} \tag{4}$$

for $j = 1, \ldots, r; t = 1, \ldots, n$ and

$$\mathbf{W}_{i,j} \leftarrow \mathbf{W}_{i,j} \frac{\sum_{t=1}^{n} \mathbf{H}_{j,t} \mathbf{V}_{i,t} / (\mathbf{WH})_{i,t}}{\sum_{t=1}^{n} \mathbf{H}_{j,t}} \tag{5}$$

for $i = 1, \ldots, m; j = 1, \ldots, r$.

Since in our scenario the spectral characteristics of the drum and harmonic sources in the signal are not known beforehand, we initialize $\mathbf{W}$ and $\mathbf{H}$ with random numbers drawn from a uniform distribution on the interval $]0, 1]$.

To reduce computational cost, instead of detecting convergence by computing the cost function (Eq. 3) after each iteration step, we run the algorithm for 100 iterations, after which in separation of popular music a reasonable separation quality is reached and convergence slows down considerably [15].

4.4 Synthesis of Harmonic Signals

Our goal is to obtain a drum-free signal from the NMF representation computed according to the previous section. To this end, we first classify the signal components $(\mathbf{w}_j, \mathbf{h}_j)$, $1 \leq j \leq r$ into two classes, 'drum' and 'harmonic'. Note that the gains $\mathbf{h}_j$ may also contain valuable features for discrimination of drum and harmonic sounds, since e. g. drum sounds are expected to be more periodic than harmonic sounds. The exact feature set and parameters used for classification will be described in the next section.

After classification, we compute a magnitude spectrogram $\mathbf{V}_{\text{harm}}$ of a signal that contains only harmonic sounds: Let $J_{\text{harm}} = \{j : (\mathbf{w}_j, \mathbf{h}_j)$ classified as harmonic$\}$. Then,

$$\mathbf{V}_{\text{harm}} = \sum_{j \in J_{\text{harm}}} \mathbf{w}_j \mathbf{h}_j \tag{6}$$

We transfer $\mathbf{V}_{\text{harm}}$ back to the time domain applying a column-wise inverse DFT, using the phase matrix from the

# beats	training	develop	test	sum
female	39 267	25 354	12 856	77 477
male	60 210	55 429	40 805	156 444
no voice	73 512	64 568	37 447	175 527
sum	**172 989**	**145 351**	**91 108**	**409 448**

Table 1: Number of beats per set of the UltraStar database.

original signal. Finally, we obtain a 'harmonic' time signal by windowing each time frame with the square root of the Hann function, then employing an overlap-add procedure.

4.5 Discrimination of Drum and Harmonic Components

For discrimination of drum and harmonic components, we use a linear SVM classifier with a feature set similar to the ones proposed in [4] and [15].

From the spectral vectors $\mathbf{w}_j$, we compute MFCC, using 26 triangular filters on a bank ranging from 20 Hz to 8 kHz. 10 first MFCC plus the zeroth (energy) coefficient are considered. Furthermore, we add sample standard deviation (using the common unbiased estimator), spectral centroid, 95 % roll-off point, and noise-likeness [19]. While these spectral features are quite common in pattern recognition, the temporal features computed from the gains vectors $\mathbf{h}_j$ are more specific to the drum beat separation task. In detail, we use percussiveness [19], periodicity, average peak length, and peak fluctuation [4, 15].

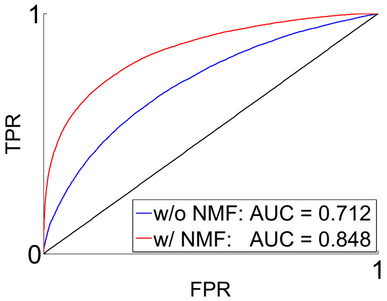
For NMF computation and extraction of the named features the open source 'Blind Source Separation for Audio Retrieval Tasks' (BliSSART)[2] package is used for reproducibility reasons. To train the SVM classifier, we used a data set generated from a popular music collection which is presented in detail in [15]. It consists of 95 drum and 249 harmonic components computed by NMF on song extracts. In 10-fold stratified cross validation of this data set, a classification accuracy of 96.2 % was achieved.

5. EXPERIMENTS

5.1 Data Partitioning

We partitioned the UltraStar database introduced in section 2 into the three groups: training, develop, and test. The training set (241 songs) contains all artists beginning with A,D,G,..., the develop set (207 songs) artists beginning with B,E,H,..., and the test set (134 artists) those beginning with C,F,I,...,0-9. Note that this dividing setup provides that all songs of one artist are in the same set, thus processing is strictly independent of the artist. See Table 1 for the gender distribution on beat level.

The features for the actual gender determination as introduced in section 3 were extracted from the original song files. In addition we created the same sets with the features extracted from the discriminated harmonic segments after applying our NMF algorithm and drum beat separation as introduced in section 4 on the original song.



(a) SVM, 2-class

Figure 1: ROC by true (TPR) over false positive rate (FPR), and the area under the curve (AUC) for the two-(female / male) class task.

We evaluate two different tasks: first, three classes (no voice / female / male) to evaluate vocal presence localization in combination with gender recognition; second, two classes (female / male) where we only consider beats with vocal presence to judge whether performance is increased particularly in gender discrimination by NMF-based drum separation.

We trained with our training set and evaluated on the develop set to verify feature relevance before and after drum-beat separation. For optimization, we applied random down-sampling – i. e. elimination of instances – to the training set to achieve a balanced distribution of instances among classes.

5.2 Results

While training with the training and evaluating with the develop set, we found that every extracted feature was relevant, as classification performance could not be improved by removing features of a certain type. We thus kept the full feature-set as described in section 3 for the oncoming evaluations. For our final results we next merged the training and develop sets for classifier training, and evaluated on the test set for representative performances.

First we consider the results with three classes on beat level (cf. Table 2, columns 'beat'): performance is improved for SVM and HNB by NMF-based drum-beat separation, and a maximum accuracy of 58.84 % is reached by HNB. Next looking at the obtained performances with only two classes (female / male) on beat level, one again notices a considerable boost by drum-beat separation for all classifiers. Again, SVM benefit most, while HNB reaches highest level at 80.22 % accuracy.

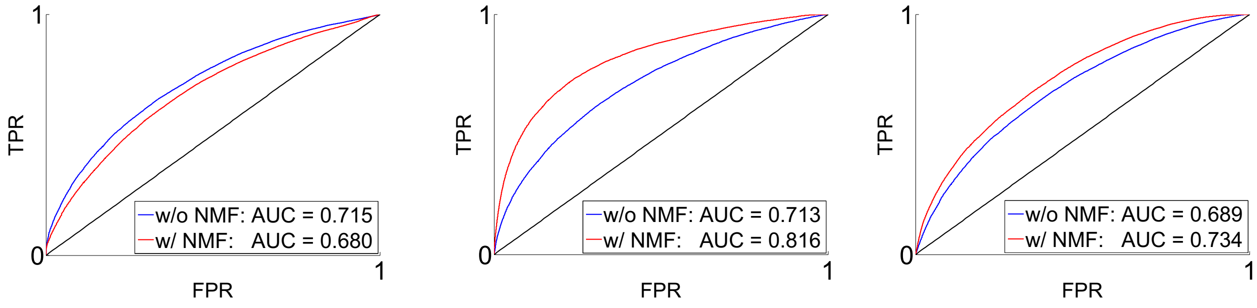
Next, we shift to the level of the whole song, and identify the gender of the mainly heard vocalist(s): after classification per beat we estimate a song's overall gender either by majority vote or alternatively by adding the prediction scores per gender and choosing the maximum overall score.

Table 2 (columns 'song') shows the according results. Minor differences – and only for HNB – are observable between majority vote and the maximum added score. The

[2] http://www.openaudio.eu

		w/o NMF		w/ NMF	
Accuracy [%]	**classification scheme**	beat	song	beat	song
-/f/m	HNB vote	58.54	79.85	58.84	84.33
	HNB added score		79.85		85.07
	SVM vote/added score	52.06	79.85	56.54	**87.31**
f/m	HNB vote	70.35	82.09	80.22	85.82
	HNB added score		83.58		86.57
	SVM vote/added score	67.52	82.09	79.97	**89.55**

Table 2: *Accuracies of vocalist gender recognition on beat level and on song level by majority voting or maximum added score for HNB and SVM as classifier – once with (w/) and once without (w/o) separation of drum-beats by NMF. Considered are no voice (-), female (f), and male (m) discrimination on all beats or only gender on those with vocal presence.*



(a) no voice (SVM, 3-class)

(b) female (SVM, 3-class)

(c) male (SVM, 3-class)

Figure 2: ROC by true (TPR) over false positive rate (FPR), and the area under the curve (AUC) for the three-class (female / male / no voice) task.

latter is found beneficial – as one would assume – as information on certainty is preserved prior to the song level decision. The results further indicate that the accuracy of classification on beat level is sufficiently above chance level to allow for repairing of mispredictions over the duration of a song. Here, SVM perform slightly better with a maximum of 87.31 % accuracy for the three classes, and the difference to the two-class task is drastically reduced. Overall, the statistical significance on song level for the improvement gained by NMF utilization is 0.05. Thus, we can state that drum separation by NMF helps recognizing gender even on the song level.

To shed light on this effect per class, according Receiver Operating Characteristics (ROC) are depicted in Figure 1 for the two-class task of gender recognition, and in Figure 2 for the three-class task with additional determination of positions that contain vocals by SVM. To provide a single value rather than a curve, one can calculate the area under the ROC curve, called AUC. The highest possible AUC is 1.0, equal to the whole graph area, and achievable only by a perfect classifier. Random guessing has an AUC of 0.5 since it corresponds to the diagonal line in the ROC space. A reasonable classifier should therefore have an AUC that is significantly greater than 0.5, with better classifiers yielding higher values. The values obtained are also shown in Figure 1: For the two-class task (female / male) the difference in the AUC with and without NMF is highly significant at the 10^{-3} level. In the three-class problem clear differences are observable: the highest benefit is reached for female vocals, next come male vocals, and interestingly the recognition of parts without vocal presence is negatively affected by reduction of the drum beat presence.

6. CONCLUSION

It was our primary goal to predict the vocalist gender in originally recorded popular music, and our secondary to analyze whether NMF usage for separation of the drum-beat can help improve on this task. The results clearly demonstrate the significant improvement obtained, and we are by that able to fully automatically identify the gender of the main vocalist in popular music at a high and reasonable accuracy for system unknown artists and songs. On beat level NMF application slightly impairs vocals presence estimation, but increases the overall performance of gender classification explaining the better results on song level.

Considering the choice of classifier, no clear tendency was found, apart from the fact that the overall best result was obtained by SVM.

Future refinement can be invested in improved annotation: as mentioned, the UltraStar annotations were created by members of the game community. Therefore errors among the ground truth tempo and vocals' locations might be present though we chose the most frequently used files. A widespread verification of the annotations would minimize the error rate and maybe reduce false classifications. But that would need a huge investment of time.

Further, we assigned the main vocalist gender. However, alternatively local labeling and consideration of mixed gender or choir passages could be provided.

Finally and self-evident, tailored vocal instead of drum separation should be targeted now that the more robustly obtainable separation of drum-beats was already found significantly beneficial.

7. ACKNOWLEDGMENT

This work was supported by the Federal Republic of Germany through the German Research Foundation (DFG) under the grant no. SCHU 2508/2-1 ("Non-Negative Matrix Factorization for Robust Feature Extraction in Speech Processing").

8. REFERENCES

[1] W. H. Abdulla and N. K. Kasabov. Improving speech recognition performance through gender separation. In *Proc. of ANNES*, pages 218–222, Dunedin, New Zealand, 2001.

[2] P. Cebula. UltraStar - PC conversion of famous karaoke game SingStar, 2009.

[3] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proc. of ISMIR*, pages 329–336, 2005.

[4] M. Helen and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. of EUSIPCO*, Antalya, Turkey, 2005.

[5] I. Kononenko. On biases in estimating multi-valued attributes. In *Proc. of IJCAI*, pages 1034–1040, Montreal, Quebec, Canada, 1995.

[6] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proc. of AAAI*, pages 223–228, San Jose, CA, USA, 1992.

[7] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proc. of NIPS*, pages 556–562, Vancouver, Canada, 2001.

[8] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. of MUSIC-IR*, Plymouth, MA, USA, 2000.

[9] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009:24 pages, 2009.

[10] A. Mesaros, T. Virtanen, and A. Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proc. of ISMIR*, pages 375–378, 2007.

[11] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, 1994.

[12] M. N. Schmidt and R. K. Olsson. Single-channel speech separation using sparse non- negative matrix factorization. In *Proc. of Interspeech*, Pittsburgh, Pennsylvania, USA, 2006.

[13] B. Schuller, F. Eyben, and G. Rigoll. Tango or Waltz?: Putting ballroom dance style into tempo detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2008(846135):12 pages, 2008.

[14] B. Schuller, C. Hage, D. Schuller, and G. Rigoll. 'Mister D.J., Cheer Me Up!': Musical and Textual Features for Automatic Mood Classification. *Journal of New Music Research*, 38:33 pages, 2009.

[15] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll. Blind enhancement of the rhythmic and harmonic sections by NMF: Does it help? In *Proc. of International Conference on Acoustics (NAG/DAGA 2009)*, pages 361–364, Rotterdam, The Netherlands, 2009.

[16] B. Schuller and F. Weninger. Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization. In *Proc. of ICASSP*, Dallas, TX, 2010.

[17] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. of WASPAA*, pages 177–180, 2003.

[18] V. Stouten, K. Demuynck, and H. van Hamme. Discovering phone patterns in spoken utterances by non-negative matrix factorization. *IEEE Signal Processing Letters*, 15:131–134, 2008.

[19] C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. of ICA*, Nara, Japan, 2003.

[20] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, March 2007.

[21] T. Virtanen, A. Mesaros, and M. Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *Proc. of SAPA*, Brisbane, Australia, 2008.

[22] T. Vogt and E. Andre. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. of LREC*, Genoa, Italy, 2006.

[23] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Proc. of DMRN Summer Conference*, Glasgow, Scotland, 2005.

[24] C. Xu, N. C. Maddage, and X. Shao. Automatic music classification and summarization. *IEEE Transactions on Speech and Audio Processing*, 13(3):441–450, 2005.

[25] H. Zhang, L. Jiang, and J. Su. Hidden naive Bayes. In *Proc. of AAAI*, pages 919–924, Pittsburgh, PA, USA, 2005.